对混合属性数据表可行的差分隐私保护方法 *

丁永善，李立新

(信息工程大学 三院，郑州 450000)

摘 要：为加强隐私保护和提高数据可用性，提出一种可对混合属性数据表执行差分隐私的数据保护方法。该方法首先采用 ICMD（insensitive clustering for mixed data）聚类算法对数据集进行聚类匿名，然后在此基础上进行 ε -差分隐私保护。ICMD 聚类算法对数据表中的分类属性和数值属性采用不同方法计算距离和质心，并引入全序函数以满足执行差分隐私的要求。通过聚类，实现了将查询敏感度由单条数据向组数据的分化，降低了信息损失和信息披露的风险。最后实验结果表明了该方法的有效性。

关键词：混合属性；聚类；差分隐私；敏感度；隐私保护

中图分类号：P309 doi: 10.3969/j.issn.1001-3695.2017.08.0729

Differential privacy protection method for mixed data

Ding Yongshan, Li Lixin

(*The* 3*th College*, *Information Engineering University*, *Zhengzhou* 450000, *China*)

Abstract: To enhance privacy protection and improve data availability, this paper proposed a differential privacy data protection method ICMD-DP for mixed data. ICMD-DP performed differential privacy on the results of ICMD (insensitive clustering method for mixed data) . To satisfy the requirement of maintaining differential privacy, ICMD used different methods to calculate the distance and centroid of categorical and numerical attributes and introduced the total order function. The combination of clustering and differential privacy realizes the differentiation of query sensitivity from single record to group record. At the meanwhile, it reduced the risk of information loss and information disclosure. Finally, this paper gave experiments to illustrate the effectiveness of the method.

Key Words: mixed data; clustering; differential privacy; sensitivity; privacy protection

0 引言

随着互联网络的发展，社交网络上的用户数量剧增，其中的用户数据和信息拥有着巨大的吸引力。数据挖掘者可以从中获取巨大价值，但同时用户隐私也面临着泄露风险[1]。为确保网络用户的隐私安全，需要对其进行保护。如何在保证用户数据可用性的前提下，最大限度保护发布的用户信息不被攻击者窃取，成为用户信息共享中的难点。

隐私保护的数据发布（PPDP）应该平衡考虑两个方面的问题:a）针对敏感信息的充分保护，消除用户数据共享时的顾虑；b）减少非敏感信息的信息损失，保证数据的最大可用性[2]。

k-anonymity 及其扩展是进行用户信息隐私保护的重要方法。*k-anonymity* 要求数据记录中至少拥有 *k* 条在准标志符上不可区分的记录，使得攻击者无法大概率地辨别出隐私信息的拥有者，从而保护了用户隐私[3]。然而，攻击者可以通过挖掘获得越来越多的背景知识，产生很多攻击变体，如背景知识攻击和同质攻击，使上述模型难以应对。攻击者通过对用户的准标志符进行组合应用，常常能够确定特定用户，进而获取该用户的其他隐私信息。因此，针对任意背景知识的隐私保护方法称为该领域研究的热点[4]。Dwork[5]提出了差分隐私保护模型并给出了严格的隐私证明，该模型克服了无法抵御任意背景知识攻击的缺点。但是差分隐私模型在保护数据隐私时牺牲了较大的数据可用性。

针对上述问题，结合聚类和差分隐私的特点，本文提出了针对混合型数据表在发布过程中的隐私保护方法，该方法假设攻击者拥有全部的背景知识，能够克服了背景知识不断扩大引起的隐私保护模型不再适用的缺点。该方法不仅满足差分隐私保护模型中的隐秘性要求，同时保证了发布数据的质量。本文的主要工作有：a）改进 MDAV，提出了针对混合属性数据表的聚类算法 CMD；b）将全序距离函数引入到 CMD，提出非敏感聚类算法 ICMD，以便更好地执行差分隐私；c）在聚类算法的基础上执行差分隐私操作，提出了 ICMD-DP 的差分隐私数据

基金项目：国家重点研发计划项目（2016YFB0501900）

作者简介：丁永善（1992-），男，河南周口人，硕士研究生，主要研究方向为信息安全（ysding200@163.com）；李立新（1967-），男，重庆人，研究员，博士，主要研究方向为网络计算、数据库、信息安全.

发布方法；d）通过实验，验证了文中所提方法在保护数据隐私和提高数据可用性方面的有效性。

1　相关工作和预备知识

1.1　相关工作

随着大数据和数据挖掘的兴起，数据隐私保护的研究也越来越得到重视。隐私保护技术大致分为噪声干扰、匿名发布和数据加密等。其中匿名技术在用户数据安全方面发挥了积极作用，如 Wang 等人[6]提出了使用匿名化技术保护用户认证信息的安全性等。本文涉及的匿名化是通过聚类算法实现的，即聚类匿名。

聚类匿名在数据挖掘和数据分组中发挥重要作用，同时也成为隐私保护研究的热点。聚类匿名的目的是将不同的对象进行划分分组，使各组之间的相似度最大，组间的相似度最小。*k*-匿名算法是一种分组算法，使得每组中的数据条数至少为 *k*；MDAV（maximum distance to average vector）算法满足计算数值属性数据表的聚类匿名化[7]。基于 *k*-匿名，文献[8]提出了一种保持网络结构稳定的 *k*-度匿名隐私保护模型；文献[9]结合贪心法和聚类划分的思想,提出一种贪心聚类匿名方法，以争取达到信息损失量和时间效率的最优化。但上述方法作为通用方法，难以应对背景知识增长的攻击变体。为更好地抵御背景知识攻击和同质攻击，保护特定的敏感值或全部敏感值，文献[10]提出了单敏感值 (α,k) -匿名模型和多敏感值 (α,k) -匿名模型；文献[11]提出了一种新的基于 $(p+,\alpha)$ -敏感 *k*-匿名隐私保护模型。但它们依然无法抵御任意背景知识攻击。Dwork[5]差分隐私保护模型的提出在一定程度上解决了上述问题；文献[12]介绍了差分隐私保护的理论基础和最新研究进展,并给出了一个差分隐私保护的应用框架 PINQ(privacy integratedqueries)。

Torra[13]针对分类型数据提出了基于 *k-modes* 的微聚集算法，但它们只针对单一的数值属性或分类属性数据进行处理。*k-prototypes* 算法通过集成 *k-means* 和 *k-modes* 算法，实现了对混合数据的聚类分析[14]，但算法参数难以确定且不能客观反映数据对象和类中心的差异性。

聚类匿名和差分隐私的结合在一定程度上解决了上述问题。其中，文献[13]提出了一种基于差分隐私保护的 *k-means* 聚类隐私保护方法，该方法是对选取的中心点和集合内点之和进行差分隐私，但该方法的聚类可用性不仅依赖于隐私保护预算，还依赖于数据集大小；文献[2]提出了基于 DBSCAN 聚类算法的差分隐私数据保护方法，但该方法只作用于数值型属性数据集。本文通过描述混合属性数据表记录之间的距离和质心计算方法，通过改进 MDAV，提出针对混合数据数据表并满足 *k*-匿名的非敏感聚类算法 ICMD(*k*-anonymity by insensitive clustering for mixed data)，并在此基础上进行差分隐私保护。

数据发布和隐私保护中，通过聚类处理，可以实现单条数据到组数据的匿名化。对聚类处理过的数据表进行差分隐私保护，可以将查询函数的敏感性进行分化，进而减少数据信息的损失，提高数据的可用性。

1.2　差分隐私保护

差分隐私保护模型是通过对原始数据集或统计结果添加噪声以达到隐私保护的目的。该模型给出了严格且强健的隐私保护证明,可以确保在数据集中更改一条记录而不影响统计结果，保证了数据集的原有统计特性。另外，该模型可抵御任意背景知识攻击[4]。

定义 1　假设数据集 D 和 D'，两者中的一个可以通过修改另一个的单一数据记录得到， $Range(A)$ 为 算法 A 的值域，若算法 A 在数据集 D 和 D' 上的任意输出结果为 $S(S\in Range(A))$ 且满足

$$Pr[A(D)=S]\le e^{\varepsilon}\times Pr[A(D')=S] \tag{1}$$

则算法 A 满足 $\varepsilon-$ 差分隐私，其中参数 ε 称为隐私保护预算， ε 越小隐私保护程度越高，同时引入的噪声越大。

差分隐私保护技术具有序列组合性和并行组合性[15]。差分隐私保护可以通过添加拉普拉斯噪声干扰查询结果而实现。

定义 2　对于查询函数 f ，若算法 A 有 $A(D)=f(D)+Lap\left(\frac{\Delta f}{\varepsilon}\right)$，则算法 A 满足 $\varepsilon-$ 差分隐私。

其中：Δf 表示查询函数的敏感性，指的是查询函数 f 作用于邻近数据集时产生的最大距离差。文献[16]给出了添加拉普拉斯噪声引起的误差 $error_{abs}^{i}=\frac{\sqrt{2}\Delta f}{\varepsilon}$ 。

1.3　混合型数据表中距离和质心计算

现有数据表大多数为混合型数据表，即表中的数据属性既有数值型又有分类型。针对不同属性的数据有不同的距离计算和质心求解方法。采用单一的方法往往会造成信息丢失、质心偏差等问题，因而本文提出一种针对混合型数据表的距离计算和质心求解方法。

设混合型数据集 D 以及 X,Y 为数据集 D 中的记录,每一个记录具有 p 维分类属性和 q 维数值属性，计算数据记录 X,Y 的距离 $D(X,Y)$ ，首先分别计算其分类属性距离 $d(X,Y)_c$ 和数值属性距离 $d(X,Y)_n$ 。定义如下。

定义 3　分类距离。 对于数据表中的任意记录 X,Y ，假设数据表含有 p 维分类属性，则记录 X,Y 的分类属性部分的距离定义为

$$d(X,Y)_c=\sum_{j=1}^{p}\delta(x_j,y_j) \tag{2}$$

其中： $\delta(x_j,y_j)=\begin{cases}0 & (x_j=y_j)\\ 1 & (x_j\ne y_j)\end{cases}$

由式(2)可知，每维分类属性取值[0,1]，对于数值属性，如果采用海明距离作为每维数据的距离，会导致分类属性部分的距离被数值属性部分的距离湮灭，因而采用如下定义计算数值属性距离。

定义 4　数值距离[17]。 首先将数据记录的数值属性部分的每一维进行标准化处理，即 X 第 q 维值为

$d(X^q)_n = \frac{X^q - X^q_{min}}{X^q_{max} - X^q_{min}}$ 其中：X^q_{max} 为该维数据记录的最大值，X^q_{min} 为该数据记录的最小值，则数值部分距离为

$$d(X,Y)_n = \sum_{j=1}^{q}(d(X^j)_n - d(Y^j)_n) \tag{3}$$

定义 5 混合距离。依据定义 3 和 4 的结果，通过把数据记录 X,Y 的分类属性和数值属性的距离相加可得它们之间的距离[18]，即 $D(X,Y)=d(X,Y)_c+d(X,Y)_n$。

定义 6 质心。设 T 是 n 维数据集 D 的一个等价类，t_i 是等价类 T 的一条记录，即 $t_i \in T,(i=1,2,\cdots,n)$，$t_i^o$ 是记录 t_i 的数值属性部分，t_i^c 是记录 t_i 的分类属性部分，即 $T=\{t_1,t_2,\cdots,t_n\}=\{\{t_1^o,t_1^c\},\{t_2^o,t_2^c\},\cdots,\{t_n^o,t_n^c\}\}$，设 t^o 是数值属性 $\{t_1^o,t_2^o,\cdots,t_n^o\}$ 的均值，t^c 是数值属性 $\{t_1^c,t_2^c,\cdots,t_n^c\}$ 的泛化，则等价类 T 的质心为 $C(T)=\{t^o,t^c\}$。

本文分别采用均值和泛化来代替等价类中的原始的数值和分类数据，避免了单一方法对数值和分类数据聚类时的片面性和误差，保留了更多的语义。

2 数据发布方法

针对混合属性数据表，阐述其距离和质心的计算方法，提出一种满足 k-匿名机制的聚类方法，然后对聚类后的数据添加噪声，实现差分隐私保护。聚类操作减小了查询函数的敏感性，进而可以通过添加较小的噪声达到同样的隐私保护效果，提高数据可用性。

2.1 对混合属性数据表可行的聚类方法

本文在 MDAV[7]的基础上，采用 1.2 节的混合属性数据表距离和质心计算方法，提出一种对混合属性数据表可行的聚类匿名化方法 CMD(clustering for mixed data)，根据 k-匿名的定义可知，该方法同时满足 k-匿名机制。

算法 1 聚类算法 $CMD(D,k)$

输入：D 为有 $n \geq 2k$ 条记录的原始数据集，k 为聚类最小尺寸。

输出：满足 k – 匿名的聚类数据集 D'。

步骤：

使用文献[19]的方法计算聚类中心，并通过定义 5 的方法计算距离该中心最远的记录 r 和距 r 最远的记录 s，作为两个初始类中心；

分别计算距离 r 和 s 最近的 k 条记录，并将其进行归类，加入到数据集 D'；

对剩下的 m 条记录，若 $m \geq 2k$，则对剩下的数据记录重复步骤 1、2；

若 $m \in [k, 2k-1]$，则自成一类，加入到数据集 D'；

否则，将剩下的 m 条记录，划归到距离各自最近的类中；

计算各类的类质心，并用其替换各类中的数据记录；

返回替换后的数据表 D'。

算法 1 返回的数据表 D'满足 k-匿名机制，其中的每个组都至少拥有 k 条记录，对每组记录中的数值属性和分类属性，分别用均值和泛化值进行替换，降低了查询函数的敏感性。

2.2 可执行差分隐私保护的聚类改进方法

差分隐私和聚类算法提供了不同的信息披露保护。利用聚类算法能降低差分隐私中需要引入的噪声，实现了查询函数的敏感性分化，同时差分隐私保护能够弥补聚类算法的不可抗任意背景知识攻击。两者的结合能够达到更好的隐私保护结果，并保留较好的数据可用性。

设 M 为聚类函数，f 为查询函数，为了有效降低 $f \circ M$ 的敏感度，M 应该满足对于数据集 D 和 D',(D 为原始数据集，D' 为对 D 修改一条记录后生成的数据集)，其聚类中心基本稳定，那么就要求数据集 D' 聚类后产生的所有簇与原本相对应的簇两两之间只有一条记录不同。本文称这样的聚类算法 M 为非敏感聚类，只有满足非敏感聚类的聚类函数才能执行差分隐私保护[20]。

定义 7 非敏感聚类。假设数据集 D，聚类函数 M，D 经 M 的聚类结果 $\{C_1,C_2,\cdots,C_n\}$，D' 为对 D 只进行修改一条记录得到的数据集，$\{C_1',C_2',\cdots,C_n'\}$ 为 D' 经 M 的聚类结果。若聚类结果 $\{C_1,C_2,\cdots,C_n\}$ 和 $\{C_1',C_2',\cdots,C_n'\}$ 对应的簇中只有一个数据记录不同，称聚类算法 M 为非敏感聚类。

为了使聚类方法 CMD 满足非敏感聚类，进而可以执行差分隐私进行数据保护，需要改变其中的距离函数 D 为一个全序函数[20]。针对混型性数据表，可通过如下方式构造满足全序关系的距离函数。

假设数据表 D 含有 n 维属性，其中 p 维分类属性，q 维数值属性，X,Y 为数据表 D 中的任意数据记录，Z 为数据表 D 的聚类中心，通过定义 5 的距离公式计算距离 Z 最远的数据记录，记为 X_b，并计算距离 X_b 最远的数据记录 X_t，定义数据表 D 的边界为[X_b, X_t]，则

$$D(X,Y) = \sqrt{\sum_{i=1}^{n}\frac{(dist(x^i,y^i))^2}{(dist(x_b^i,x_t^i))^2}} \tag{4}$$

是满足全序关系的距离函数。

其中：$dist(x^i,y^i) = \begin{cases} \delta(x^i,y^i) & 数据表D的第i维为分类属性 \\ |x^i\text{-}y^i| & 数据表D的第i维为数值属性 \end{cases}$。

将上述距离函数引入聚类算法 CMD，构造满足非敏感聚类的聚类算法 ICMD(insensitive CMD)。

算法 2 非敏感聚类算法 $ICMD(D,k)$

输入：D 为有 $n \geq 2k$ 条记录的原始数据集，k 为聚类最小尺寸。

输出：可执行差分隐私保护的聚类数据集 D'。

步骤：

计算原始数据集的边界 $[X_b, X_t]$；

分别计算距离 X_b 和 X_t 最近的 k 条记录，并将其进行归类，加入到数据集 D'；

对剩下的 m 条记录，若 $m \geq 2k$，则对剩下的数据记录重复步骤 2；

若 $m \in [k, 2k-1]$，则自成一类，加入到数据集 D'；

否则，将剩下的 m 条记录，划归到距离各自最近的类中；

计算各类的类质心，并用其替换各类中的数据记录；

返回替换后的数据表 D'。

对算法 2 中的距离计算采用式（4）的计算方法，则 ICMD 满足非敏感聚类算法定义，可对其结果执行差分隐私保护。由文献[20]可知，对于查询函数 f_i（返回数据集中的第 i 条记录），有 $\Delta(f_i \circ ICMD) \le \frac{\Delta(f_i)}{k}$。由此可知，原始数据集经过聚类分组，实现了记录隐藏和查询敏感性由单条数据向组数据的分化。

2.3 差分隐私保护数据发布方法

基于 k-匿名机制的聚类匿名不能够抵御背景知识攻击和同质攻击，为了进一步保护，在聚类的基础上对数据记录添加噪声，以达到差分隐私保护的目的。采用文献[20]的方法添加拉普拉斯噪声，实现一种对混合属性数据表实施噪声扰动的数据保护方法 ICMD-DP(clustering for mixed data with differential privacy)。

算法 3　差分隐私保护算法 $ICMD-DP(D,\varepsilon)$

输入：D 为有 $n \ge 2k$ 条记录的原始数据集，ε 为隐私保护预算。

输出：满足 $k-$匿名的 $\varepsilon-$差分隐私数据集 D_ε。

步骤：

对数据集 D 进行聚类处理 $ICMD(D,k)$，返回数据集 D'；

查询函数 f_i 返回数据集 D' 第 i 条记录的属性，函数 $S_\varepsilon()$ 为查询结果添加拉普拉斯噪声。则对于 $i \in (1,n)$，$x_i = S_\varepsilon(f_i(D'))$，将 x_i 加入数据集 D_ε；

返回数据集 D_ε。

每个查询函数的结果满足 $\varepsilon-$差分隐私，又每条查询针对的记录不相交，则根据并行性原则[15]可知，最终的数据集 D_ε 满足 $\varepsilon-$差分隐私。

对于聚集尺寸为 k 的数据集 D，单个查询敏感度小于 $\Delta f_i(D)/k$，并且有 n/k 个相互独立的查询，因此若要满足经 $ICMD-DP$ 差分隐私保护的数据查询敏感度小于原始数据集的查询敏感度，则需有 $\frac{\Delta f_i(D)}{k} \times \frac{n}{k} < \Delta f_i(D)$，即 $k > \sqrt{n}$。由上可知，虽然经聚类算法处理将造成信息丢失，但该部分损失可以由敏感度降低带来的增益进行弥补。

3 实验

本章将从时间消耗、信息损失和泄密风险等方面对文中提出的方法进行实验分析。

3.1 实验数据和环境

本文实验数据采用 UCI 的 Adult 数据集（http://archive.ics.uci.edu/ml/datasets/Adult），该数据集常用来评估隐私保护方法。该数据集为混合属性数据集，包含 6 个数值属性（如 age、hours-per-week 等）和 8 个分类属性（如 occupation、native-country 等），该数据集共有 48 842 条数据记录，选取其中不包含空值的 30 000 条数据记录进行实验。

3.2 实验方法

采用和 k-匿名评估类似的方法对本文提出的结合方案进行评估，包括信息损失（影响数据可用性）和信息披露（揭示隐私保护程度）两个方面。

3.2.1 信息损失

信息损失是指匿名数据集和原始数据集之间的差异，通常用 SSE(sum of squared errors)进行度量[20]。SSE 表示了匿名数据集和原始数据集中所有记录的属性距离的平方和，即

$$SSE = \sum_{x_j \in X} \sum_{a_j^i \in x_j} (dist(a_j^i,(a_j^i)'))^2$$

其中：a_j^i 是原始数据集中第 j 个记录的第 i 个属性，$(a_j^i)'$ 是匿名数据集中第 j 个记录的第 i 个属性。对于分类属性和数值属性，距离函数 $dist()$ 分别采用式（2）和（3）进行计算。SSE 的值越大，信息损失越严重，数据的可用性越差。

3.2.2 信息披露

信息披露通过使用匿名数据集中的记录成功匹配到原始数据集中记录的概率进行度量，又称为记录关联（record linkages ,RL）。

$$RL = 100 \times \frac{\sum_{x_j \in X} \Pr(x_j')}{n} \tag{5}$$

其中：n 是数据集中记录的个数。式（5）中的 $\Pr(x_j')$ 的计算方法如下：

$$\Pr(x_j') = \begin{cases} 0 & 如果 x_j \in G \\ \frac{1}{|G|} & 如果 x_j \notin G \end{cases}$$

其中：G 是数据记录 x_j' 所在的集合，如果记录 x_j 也在集合 G 中，就认为有概率 $\frac{1}{|G|}$ 造成信息披露；否则，该概率为 0。

为了更好地说明本文方法的有效性，计算聚类算法 CMD 和标准的 $\varepsilon-$差分隐私的 SSE 和 RL 作为基础，同时分别计算 ICMD 以及 ICMD-DP 的 SSE 和 RL 进行对比。另外 ε 采用常用取值 0.01、0.1、1、10，k 取值 2~500。

3.2.3 实验结果和分析

分别以聚类算法 CMD 和标准 $\varepsilon-$差分隐私算法为基准，通过调整 k 的取值，做对比实验，结果如图 1 所示。

由图 1、2 可知，非敏感聚类算法 ICMD 比原始聚类算法 CMD 造成更大的信息损失，但也相应地降低了信息披露的风险；在非敏感聚类算法 ICMD 的基础上进行差分隐私，能有效降低信息披露风险，起到更好的隐私保护效果，且隐私保护预算越小，隐私保护效果越明显，但同时也造成了更大的信息损失。由图 3 可知，随着 k 值的增大经算法聚类后的差分隐私的信息损失逐渐减少，且在 $k = \sqrt{n}$ 附近时和标准差分隐私具有相似的信息损失，但当 $k > \sqrt{n}$ 时，其信息损失量逐渐比标准差分

隐私更小。由图 4 可知，经聚类后的差分隐私的信息披露更小，其具有比标准差分隐私具有更好的数据保护效果。

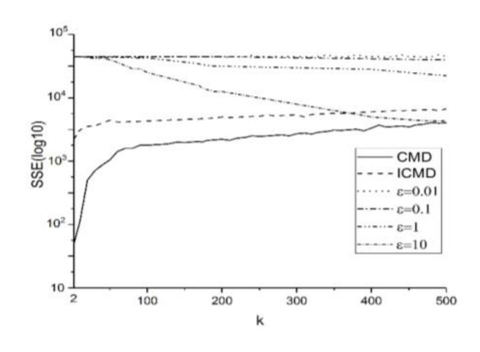


图 1　不同隐私保护预算下 ICMD-DP 与 CMD、ICMD 的信息损失量对比

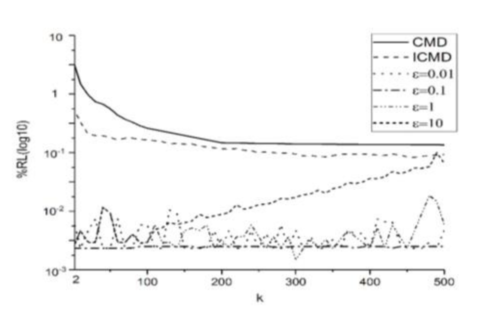


图 2　不同隐私保护预算下 ICMD-DP 与 CMD、ICMD 的信息披露对比

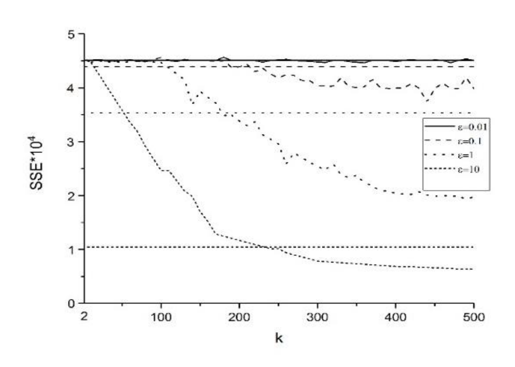


图 3　不同隐私保护预算下 ICMD-DP 与标准差分隐私保护的信息损失量对比

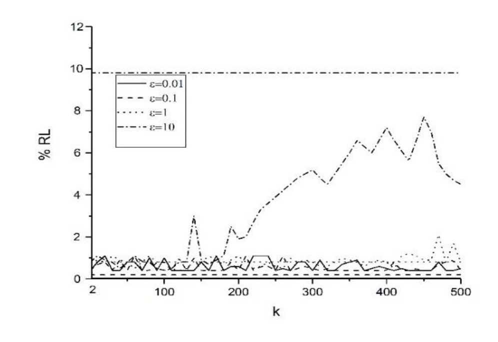


图 4　不同隐私保护预算下 ICMD-DP 与标准差分隐私保护的信息披露对比

4　结束语

本文提出一种针对混合属性数据表可行的数据发布方法 ICMD-DP，该方法将聚类和差分隐私相结合，平衡数据可用性和隐私保护之间的矛盾。ICMD-DP 通过聚类匿名后执行差分隐私，降低了信息披露的风险，同时增加了数据可用性。首先，在 MDAV 的基础上提出了对于混合数据表的聚类算法 CMD；其次，为了更好地执行差分隐私，将全序距离函数引入到 CMD，提出非敏感聚类算法 ICMD 并将 ICMD 的执行结果作为输入，执行差分隐私进行数据保护；最后，通过实验分析了该方法在混合属性数据表上保护用户隐私和提高数据可用性上的有效性。

参考文献：

[1] 曹春萍，郑夏．社交网络中隐私保护的匿名模型研究 [J]．小型微型计算机系统, 2016, 37 (8): 1821-1825.

[2] 刘晓迁，李千目．基于聚类匿名化的差分隐私保护数据发布方法 [J]．通信学报, 2016, 37 (5): 125-129.

[3] 刘向宇，王斌，杨晓春．社会网络数据发布隐私保护技术综述．软件学报, 2014, 25 (3): 576-590.

[4] 李洪成，吴晓平，陈燕．MapReduce 框架下支持差分隐私保护的 K-means 聚类方法 [J]．通信学报, 2016, 37 (2): 124-130.

[5] Dwork C. Differential privacy [J]. Lecture Notes in Computer Science, 2006, 26 (2): 1-12.

[6] Wang D, He D, Wang P, et al. Anonymous two-factor authentication in distributed systems: certain goals are beyond attainment [J]. IEEE Trans on Dependable & Secure Computing, 2015, 12 (4): 428-442.

[7] Domingo-Ferrer J, Sánchez D, Hajian S. Database privacy [M]. [S. l.] : Springer International Publishing, 2015.

[8] 龚卫华，兰雪锋，裴小兵，等．基于 K-度匿名的社会网络隐私保护方法 [J]．电子学报, 2016, 44 (6): 1437-1444.

[9] 姜火文，曾国荪，马海英．面向表数据发布隐私保护的贪心聚类匿名方法．软件学报, 2017, 28 (2): 341-351.

[10] 杨高明，杨静，张健沛．聚类的（α, k）-匿名数据发布 [J]．电子学报, 2011, 39 (8): 1941-1946.

[11] 黄石平，顾金媛．一种基于（p+, α）-敏感 k-匿名的增强隐私保护模型 [J]．计算机应用研究, 2014, 31 (11): 3465-3468.

[12] 李杨，温雯，谢光强．差分隐私保护研究综述 [J]．计算机应用研究, 2012, 29 (9): 3201-3205, 3211.

[13] Torra V. Microaggregation for categorical variables: a median based approach [C]// Proc of Privacy in Statistical Database. 2004: 162-174.

[14] 赵兴旺，梁吉业．一种基于信息熵的混合数据属性加权聚类算法 [J]．计算机研究与发展, 2016, 53 (5): 1018-1028.

[15] 李杨，郝志峰，温雯，等．差分隐私保护 K-means 聚类方法研究 [J]．计算机科学, 2013, 40 (3): 287-290.

[16] 张啸剑，孟小峰．面向数据发布和分析的差分隐私保护 [J]．计算机学报, 2014, 37 (4): 927-949.

[17] 熊平，朱天清，王晓峰. 差分隐私保护及其应用 [J]. 计算机学报, 2014, 37 (1): 101-122.

[18] 夏赞珠. 微数据发布中的隐私保护匿名化算法研究 [D]. 金华：浙江师范大学, 2011.

[19] 张慧哲，王坚. 基于初始聚类中心选取的改进 FCM 聚类算法 [J]. 计算机科学, 2009, 36 (6): 206-209.

[20] Soria C J, Domingo F J, Nchez D, et al. Enhancing data utility in differential privacy via microaggregation-based k-anonymity [J]. Vldb Journal, 2014, 23 (5): 771-794.